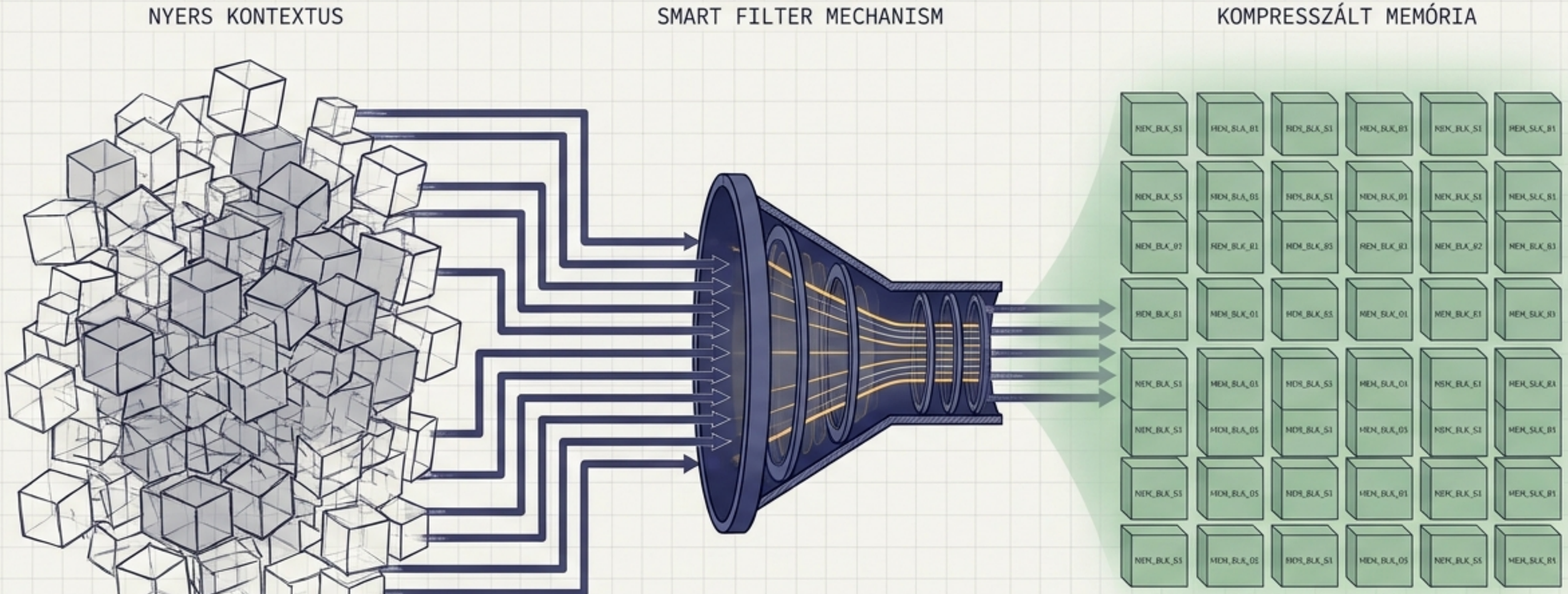


Billion Context: Végtelen memória AI kódoló ügynököknek

Hogyan küszöböljük ki a kontextus ablak korlátait a DeepSeek és a Cline használatával?

Gemini Notebook

A probléma

A munkamenet hossza növekszik.

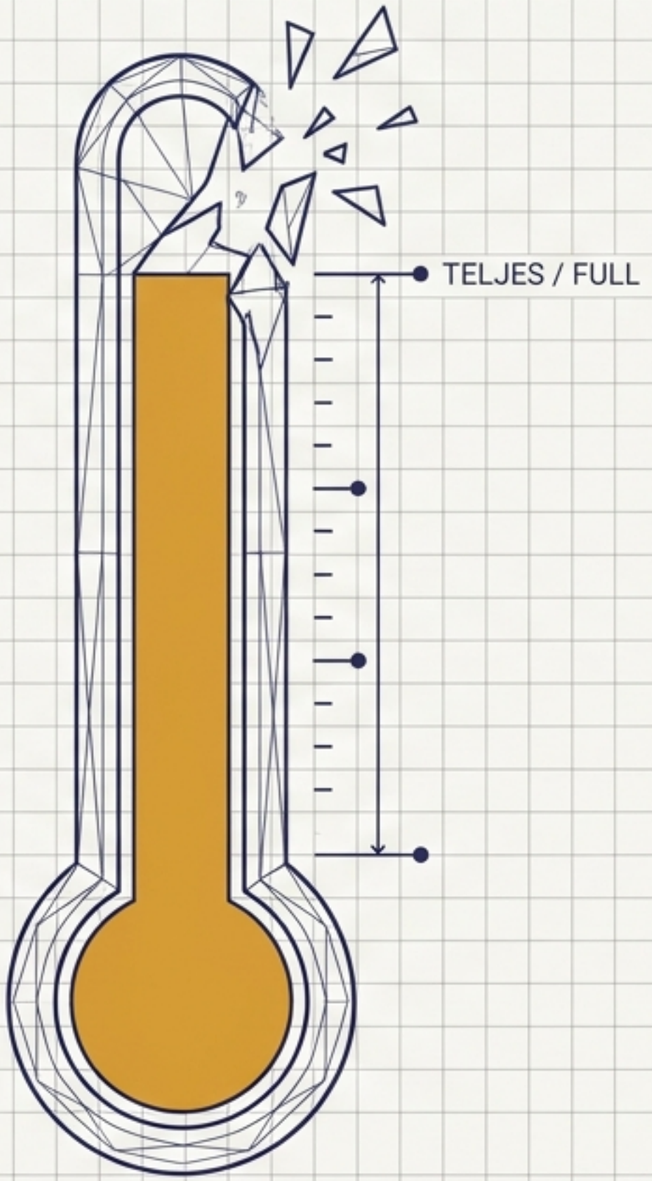
A tünet

A **kontextus ablak megtelik**. Az ügynök egyetlen elmosódott összefoglalóba tömörít mindent, vagy egyszerűen összeomlik.

A következmény

Minden egyes lépésnél a teljes történetet újra elküldi.

Eredmény: Ugyanazokért a tokenekért fizetsz újra és újra.



Gemini Notebook

A megoldás: Egy intelligens Proxy réteg

GitHub: ~3,700 Star | 300 Fork | Kevesebb mint 3 hónap alatt



- Hónapokig tartó, egyetlen munkamenet.
- Milliárdnyi token egyetlen kontextus ablakon keresztül.
- Egy 100k-s ablak elegendő.
- Ötször kevesebb felhasznált token.

Gemini Notebook

A modell saját magát menedzseli (A 4 új AI eszköz)

`compress`

Üzenetek egy tartományát részletes összefoglalóvá tömöríti.

`decompress`

Visszaállítja a pontos részleteket, ha azokra újra szükség van.



`search_context`

Kulcsszavas keresés az összes tömörített adatban.

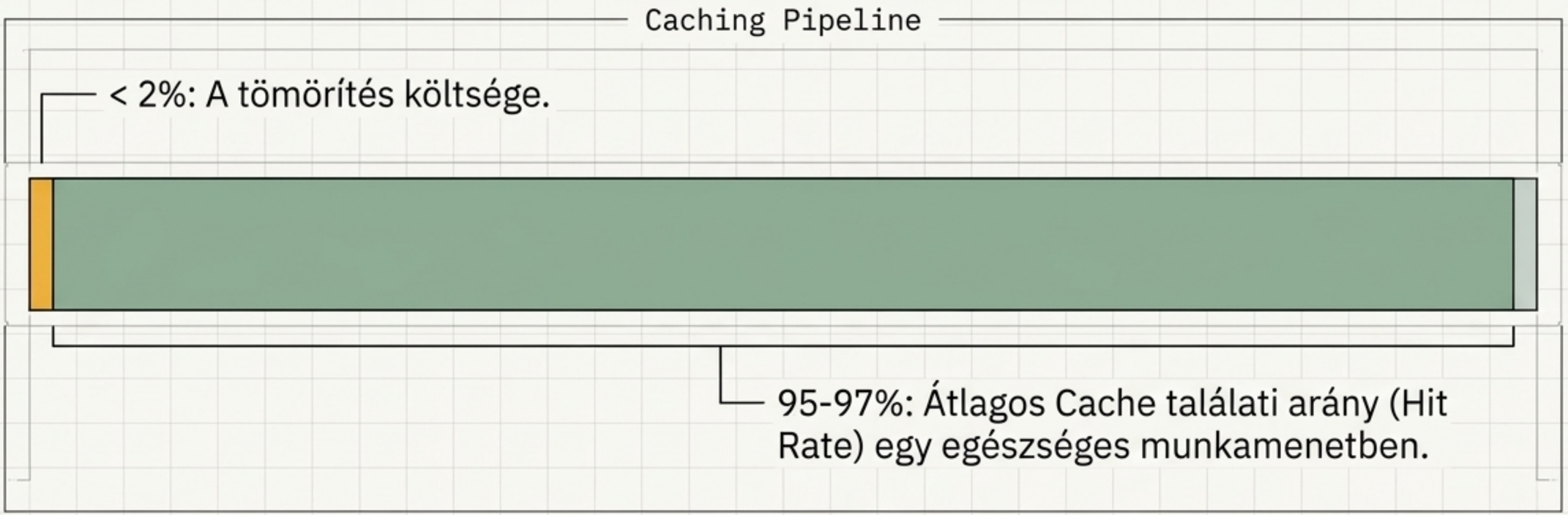
`ACP_status`

Megmutatja, mennyire van tele a kontextus, és mi tömöríthető még.

Gemini Notebook

A Prompt Cache védelme

Miért jobb ez, mint az ügynökök beépített összefoglalója?



Mivel a cache-elt tokenek rendkívül olcsók, a folyamatos, inkrementális tömörítés radikálisan csökkenti az API költségeket.

Gemini Notebook

Benchmark Analízis (DeepSeek v4.1 Flash környezetben)

Metrika	Hagyományos (Full Context)	Billion Context	Legjobb RAG
Kódolási F1 Pontszám (Contextbench)	0.29	**0.31**	-
Költség (Ugyanazon a feladaton)	$11.50	**$4.61 (58% olcsóbb)**	-
Hosszú memóriájú kiértékelés (Long memoval)	83.2%	79.2%	62.6% (71%-kal magasabb költség mellett)

Egy 4,5 hónapos, valós teszt során (174.000 API hívás, 18,76 milliárd bemeneti token) a 200k-s modelleken a kontextus ablak túlcsordulása 0 alkalommal fordult elő.

Gemini Notebook

A kompromisszum: Túlélés vs. Szó szerinti pontosság



A 1,48 milliós token teszt

- Hagyományos kontextus: 199/199 alkalommal elbukott.
- Billion Context: 199/199 alkalommal túlélte (~30x túélte (~30x tömörítéssel).



Tű a szénakazalban teszt

- A pontos, szó szerinti mondatok visszakeresési aránya (Verbatim Recall) 69%-ról 27%-ra esett.
- Tanulság: Az összefoglalók megtartják a jelentést, de nem minden pontos szót. **Kódoló ügynökök számára ez a megfelelő kompromisszum.**

Gemini Notebook

Három telepítési és futtatási stratégia

```
npm install -g billion-context --prefix <local>
```

Natív Plugin

`billy plugin install <client>`

(PI Open Code, Kimmy Code, Hermes)

Közvetlen Indítás (Launcher)

`billy <client>`

(Elindítja a klienst a proxyn keresztül a konfiguráció módosítása nélkül)

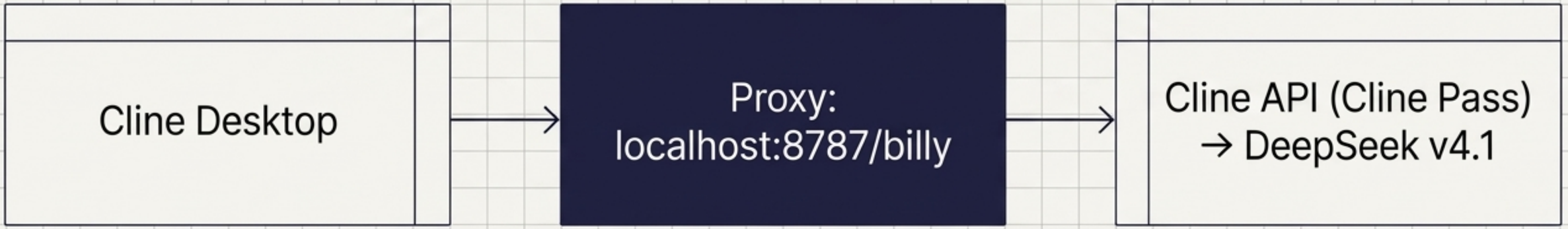
Önálló Proxy (A legrugalmasabb)

Csak elindítod a billy-t.

A proxy címe (**localhost:8787/billy**) adható meg Base URL-ként bármely kliensben.

Gemini Notebook

Integrációs Blueprint: Cline + Cline Pass + DeepSeek




```
Provider: OpenAI Compatible

Base URL: localhost:8787/billy/ + <cline-api-address>

API Key: A te Cline API kulcsod

Modell: clinepass/deepseek-v4.1-flash
```

Miért működik ez jól?

A Cline Pass fix áras, de heti/havi korlátai vannak. A Billion Context radikálisan lecsökkenti a felhasznált tokeneket, így a kvóta kitart egy hosszú munkamenet során.

Gemini Notebook

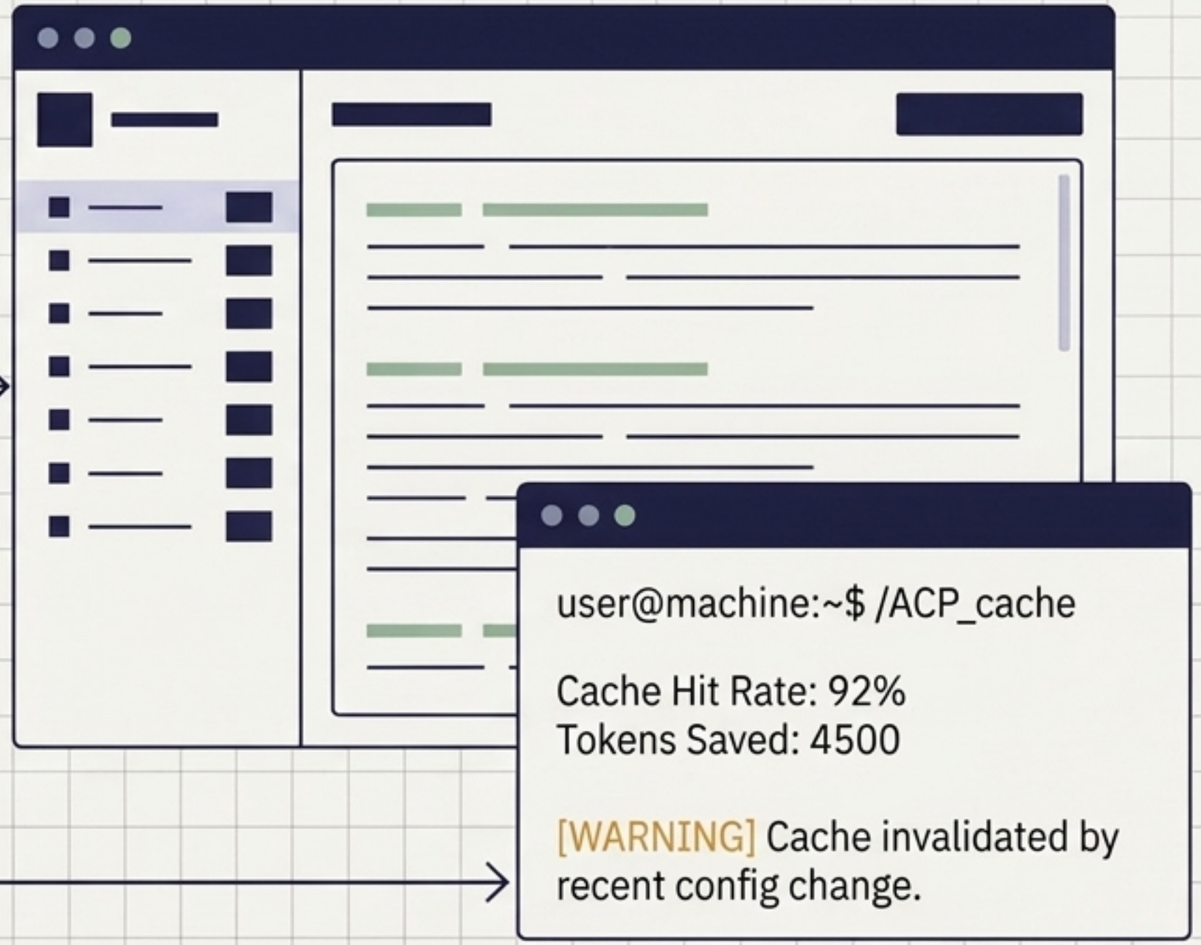
Vizuális Monitorozás és Optimalizálás

Web Dashboard (localhost:8787)

- Élő munkamenetek listája
- Valós idejű logok és konfigurációs szerkesztő

Konzol Parancs (/ACP_cache)

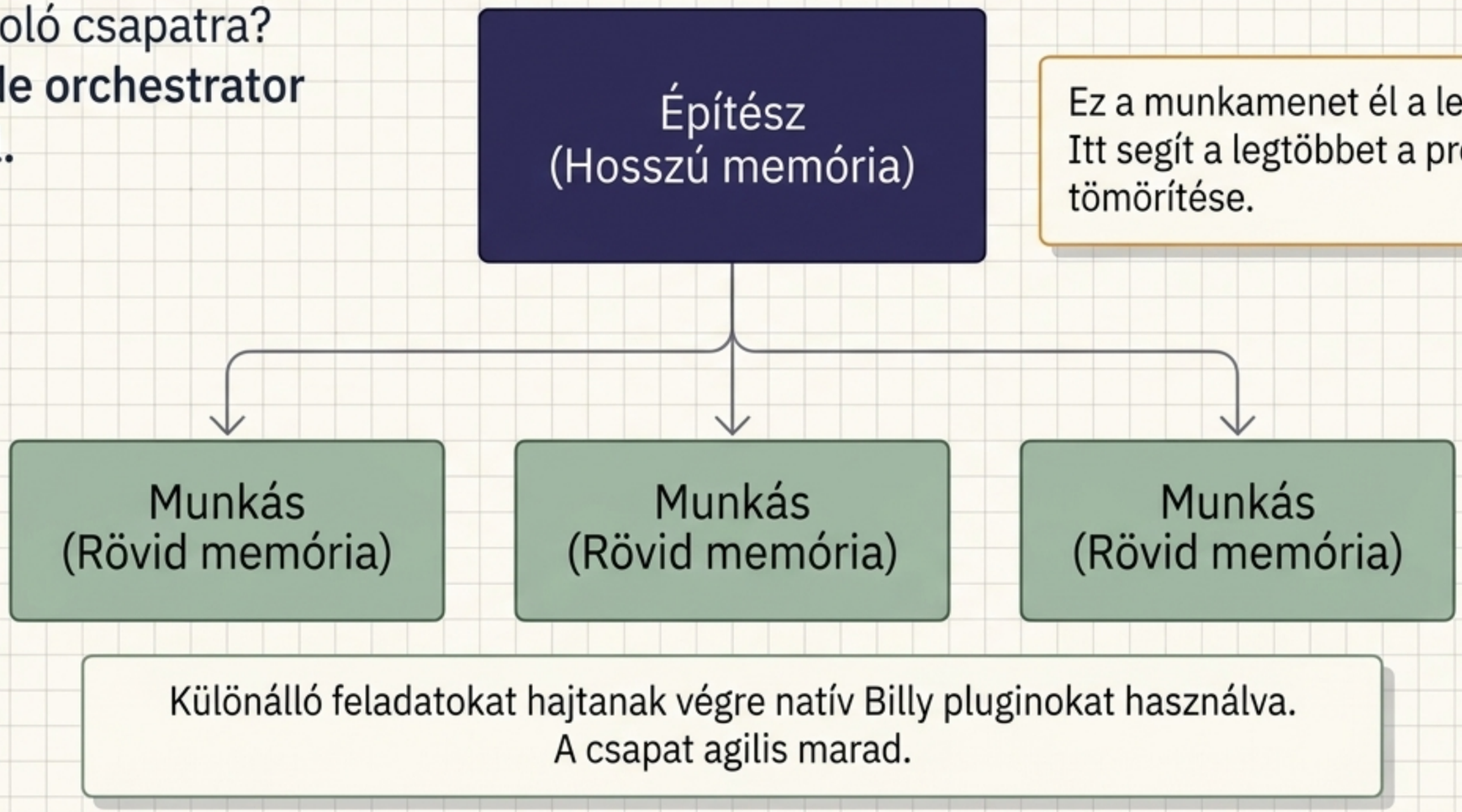
- Jelentés a pontos Cache Hit Rate-ről.
- Megmutatja a megtakarított tokeneket.
- Jelzi, ha valami megtörte a cache-t.



Gemini Notebook

Mester Architektúra: Panda Code Szintézis

Hogyan méretezzük ezt egy teljes AI kódoló csapatra?
A Panda Code orchestrator segítségével.



Gemini Notebook

Korlátok és Biztonsági Kockázatok

1. Verzió stabilitás

Gyorsan fejlődik. A `master` ág eltörhet. Mindig az `npm` release verziót telepítsd.

2. Biztonság

Hálózati futtatásnál (`--host) nincs hitelesítés (auth). Csak megbízható hálózaton használd.

3. Kompatibilitás

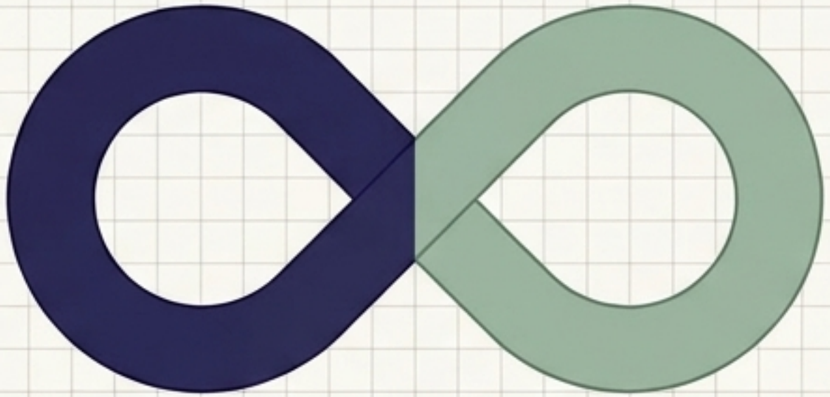
A Claude asztali alkalmazás ‘Code’ fül funkciója megkerüli a proxyt (nincs tömörítés).

4. Dokumentáció

Hatalmas és technikai, a beállítás igényel némi olvasást.

Gemini Notebook

Összegzés: Munkamenetek, amik sosem halnak meg



- **Az Architektúra:** Egy proxy, amely a 4 új eszközzel képessé teszi a modellt saját történetének inkrementális tömörítésére, a cache megőrzésével.
- **A Bizonyíték:** Olcsóbb és hatékonyabb, mint a Full Context, kiemelkedő eredménnyel a `DeepSeek v4.1 Flash` modellen.
- **A Megvalósítás:** `Cline Desktop` + `Cline Pass` + `Billion Context`.

Ha a hosszú ügynök-munkameneteid folyamatosan megszakadnak vagy túl sokba kerülnek, ez a proxy réteg az a paradigma váltás, amire szükséged van.

Gemini Notebook